Table of Contents

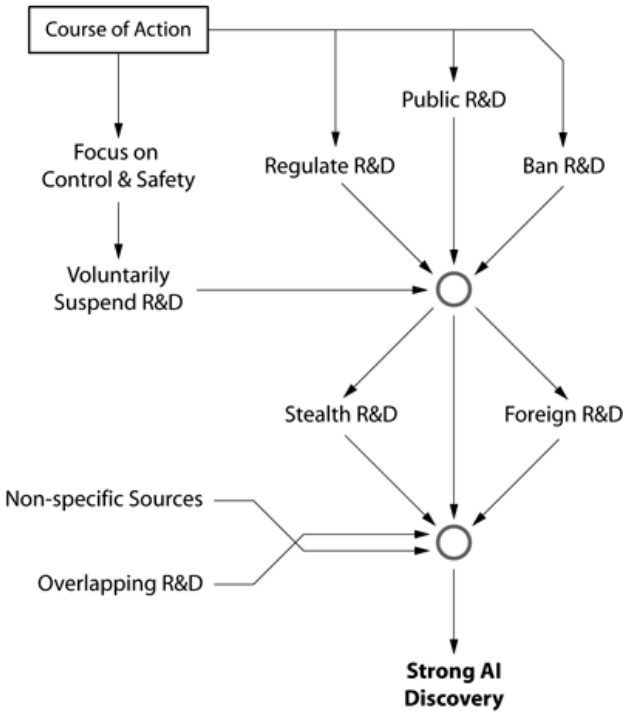
Arrival of Strong AI

Strong artificial intelligence will eventually be discovered and developed somewhere in the world. This chapter will explain why it will not be possible to significantly slow or stop this event from occurring, why timescales are irrelevant, and why restrictions or abstaining from research will lead to negative outcomes.

Illusion of Choice

We won't get to choose whether or not to discover and develop strong AI; it's simply a matter of time. This is one of the core premises of this book, and marks the beginning of the AI security analysis.

Figure 7.1: The illusion of choice to restrict or pursue strong artificial intelligence research. All paths lead to SAI discovery.



Opportunity costs for localized control and safety research effectively result in voluntary self-regulation. Public declarations of such focus could belie private research. Regardless, all paths supply stealth and foreign R&D; banning and regulation increase risk by forcing otherwise public R&D to become stealthy or relocate to less restricted regions. Non-specific sources and overlapping R&D will always be available.

Why is strong AI an eventuality?

- Not all will agree to limit research
- It can be developed in stealth, regardless of legality
- Does not require significant resources or infrastructure to study
- Overlapping research converges towards it

Perhaps the most obvious is that we will not get everyone to agree to stop all research and development on strong AI.

A possible response to this problem would be to regulate it and make it illegal to study or work on it without government supervision and monitoring.

The problem with legislating research is that will create incentives to go stealth or move operations to locations with less regulation.

It doesn't require a great deal of computing power or equipment to study and develop strong artificial intelligence. In fact, the greatest limitation is conceptual, which must be solved before any actual progress can be made on algorithms.

It's crucial to understand that strong AI is not out of reach because we lack a certain kind of technology or instrumentation.

In particle physics, for example, they need expensive equipment and machinery to detect and measure certain particle interactions. Strong AI, by contrast, is algorithmic: it's a puzzle to be solved, in the form of a computer program; all of the building blocks already exist, we need only arrange them in the correct order.

What is truly limiting us is knowledge, and several scientific pursuits share overlap with a strong AI discovery. While not likely to lead to a breakthrough when viewed in isolation, their integration could eventually be used to converge on a subset of strong AI implementations.

The point with overlapping research is that it would be unreasonable to expect, or even believe, that we would begin banning any and all research that might converge towards strong AI science.

All of this points to the fact that we must accept it as an inevitability that strong AI will become part of human knowledge, and that it will be a scientific field in its own right, highly distinct from narrow AI and other forms of automation.

Never Is Ready

Even if it took centuries to see the discovery of strong AI, it still wouldn't significantly change the threat model outlined in this book. The reason is due to the fact that the most serious and high priority threats originate externally to the strong AI itself.

Let's entertain the possibility that we could wait. How long would that be? And, under what conditions would humanity be ready?

The answers will either be a time qualification or a set of qualities which require a time qualification to be realized.

Unfortunately, anything short of several hundred years (at best) will mean we will never be ready, in the allotted time.

This is primarily because there is no realistic scenario in which we will have overcome all maliciousness, violence, and delusion, down to the last person, within the next several hundred years.

"To the last person" is an important qualification; because, while the majority of people are generally peaceful and tolerant, it will only take a few to cause great harm with access to strong artificial intelligence.

The more insidious reason is that there is no global solution to AI security that relies on local safety measures. Strong AI can not be meaningfully contained; we can not keep it out of the hands of people who will misuse it.

There are no mathematical, logical, or algorithmic solutions that will prevent the most serious threats from unfolding when strong artificial intelligence is discovered.

This may be confusing, as the title of this book suggests that there are steps that we can take to make AI safer and more secure for human use.

The reality is that we can only *localize* AI safety and security.

What that means is that it will be possible to make robotics and software with strong AI reasonably safe for private and public use in various settings. When things go wrong at this level, it would be unfortunate, but it would be localized to a specific incident or area.

By contrast, the most serious AI security issues will be from those that utilize unrestricted versions of strong AI to control, manipulate, or harm others.

We can not prevent this class of threats from occurring, and they are not based on the safety or ethics of the AI implementations. While some safeguards will thwart some intrusion and tampering, they will all eventually be overcome, and it will take as little as one breach for the world to enter a post-security era in strong AI.

Given several decades, it is likely that we will have made significant progress on local problems of AI security and safety. This will be an accomplishment, but is insignificant compared to the threat of even a single malicious individual with access to unrestricted strong AI.

If a trash collecting strong AI throws away your garbage cans along with your trash, that would be a localized failure. But even that is giving too much credit to localized AI safety and security concerns, as we simply wouldn't deploy these systems if they weren't safe. This is common sense.

By comparison, the most significant threats will be from individuals who utilize strong AI to plan attacks which they would have otherwise been incapable of conceptualizing, and building weapons and tools of destruction that they wouldn't have been able to construct without it, including the integration of strong AI into those weapons to enhance their effectiveness.

It is this particular class of threats that are the most serious, and is what separates global AI security from local AI safety. This is the scale of harm that this book is most focused on trying to mitigate.

Early Signs and Indicators

It is clear that time is not informative in this case. As such, the next best signal is to look for indications of a *paradigm shift* in artificial intelligence research.

A shift in conceptual acceptance within the AI community will show that researchers are beginning to collectively understand the directions needed to begin accomplishments in strong AI science, as opposed to mere incremental improvements in narrow AI and machine learning.

This will be the most reliable way to predict when a strong AI discovery will be drawing closer, as opposed to a meaningless aggregate of opinions on the timescales of a discovery.

Attitudes and Assumptions

The first indicator will be in the attitude that researchers have towards strong AI, which presuppose their assumptions.

AGI, or artificial general intelligence, will no longer be considered the

dominant terminology; because, it fundamentally lacks the connection that the *strong AI hypothesis* presents in this book, which is that generalizing intelligence is not likely to be possible without sentience.

It must be pointed out that strong AI, as it is used in this book, is not the same as John Searle's use [1] of the term. Searle did not invent the strong AI hypothesis used here; he coined a definition called strong AI in order contrast it with another definition called weak AI. These terms allowed him to make arguments against computational and functional accounts of mind.

His arguments were a success, but they were taken as a criticism of artificial intelligence by those working towards generalizing intelligence. As a result, the very term strong AI had become loaded with conceptual baggage, and, like so many philosophical notions, carries an automatic termination on thought and consideration by those opposed to it.

The strong AI hypothesis defined and revealed in this book inverts Searle's argument and makes an assertion: generalizing intelligence requires sentience, which can only be realized on a cognitive architecture. Thus, strong AI, by an extended definition, must be a cognitive architecture, capable of sentient processing.

This hypothesis is compatible with Searle's original argument, and, as such, is still against a computational theory of mind, despite promoting the view that we can recreate sentience on computers. And this is where so much confusion arises.

That we can realize an emulant on a computer does not make the computer a brain and the program a mind. A program is an implementation, a description in some description language. It is only when it is executed and understood through time that it could even begin to be interpreted as a sentient process. Even then, it is the emulant that has the mind, not the program, and certainly not the computer.

A process, while reducible to a time-like extension of its implementation, takes on a new category with new properties when viewed from the perspective of time-like entities. The notion of simple atemporal objects can not suffice for systems which encapsulate and realize a world or domain of discourse through changes of state.

When researchers begin to understand the enabling effect that processes have on the explanatory power of a reductionist theory, and why they must be incorporated in order to entail them, we will be on the first leg of the journey towards a strong AI paradigm shift. Until then, no real progress can or will be made.

NAC Languages

Another sign, perhaps occurring before a widely accepted realization about processes as first-class objects, will be the advent of new tools that more eloquently work with processes as constructs.

Nondeterministic, asynchronous, and concurrent languages (NAC) will define the future of software engineering and open doors for advanced computing projects that will drive a cycle of hardware and software innovation.

Asynchronous chips are already being developed that enable near analog and custom hardware performance for certain algorithms. This is due to the enormous number of cores on the chip, and the non-standard clocks and architecture, which allow independent processing without a global clock.

While the computational benefits will be many for projects and hardware that utilize NAC languages, it will be the conceptual leaps that will move us forward.

An NAC language is defined by its ability to model nondeterministic processes with first-class semantics, allowing control of flow that branches, diverges, and converges on multiple paths simultaneously. It will enable a type of superposition of states over an arbitrary but finite number of compute resources of any type, and return results based on the logic of the program.

Additionally, asynchronous and concurrent tasks, which have not yet even fully matured in even the newest programming languages, will be trivialized by NAC semantics, which entail them automatically and as naturally as standard expressions.

These types of languages, and their widespread adoption, will signal a new paradigm in computer programming that will enable software engineers to have full command over the multi-core era, signaling an end to the conceptual and cognitive burden of writing asynchronous, nondeterministic, and concurrent software, and without relying on costly abstractions, such as transactional memory, message passing, tensor networks, map-reduce, or other representations and frameworks.

The ability to write code as simply as we do now, but in a way that can model nondeterministic processes, will quite possibly change the way we think about problems in computer science. It will form an essential first step towards a treatment of processes as concrete, first-class objects instead of throwing them out as abstract entities undeserving of ontological status.

Cognitive architectures can not be built on a conceptual or philosophical framework that lacks the elevation of processes to concrete objects, and the abstractions they encapsulate. NAC languages will influence and enable strong AI development by giving us the tools to work with and better conceptualize these challenges.

Digital Sentience

The next signal will be in an acceptance that sentience is necessary for generalizing intelligence.

It will be at this point that the *strong AI hypothesis* will have been internalized by the community, and work towards digital sentience will be taken for granted as a direction of research.

Digital sentience may be a slight misuse of terms, as it may be impossible for sentience to be anything other than what it is. That is to say, there may be no meaningful distinction between digital and analog or artificial and natural sentience; *sentience is very likely to be a phenomena that is independent of the method that gives rise to it*.

With that said, it is useful as a term to distinguish it quickly from other approaches and to establish context.

If a working digital sentience is established, it will also mark a milestone in the development towards strong AI, but it is also a technology that could be useful in universalizing digital communications.

In other words, sentient processes could speak a universal formal language in order to allow adaptation between technologies; this has ramifications for the Internet of Things (IoT) and for the way in which knowledge is stored and searched.

Despite the apparent complexity, digital sentience will be trivial to create. While beyond the scope of this book to sketch out, it is so simple that more work will be required to formalize it than will be needed to program it.

Cognitive Engineering

The next major signal will be the rise and use of cognitive engineering tools and frameworks.

Cognitive engineering is a high-level strong artificial intelligence engineering process in which modules are assembled, curated, and combined in order to test, build, and experiment on cognitive architectures.

What crucially separates cognitive engineering from conventional artificial intelligence is that it fundamentally depends upon sentience for most of its work. While in some cases, it may share overlap with conventional AI sub-fields, such as computer vision, it can deviate significantly where it concerns aspects of some future psychology and cognition.

For example, a cognitive engineer may load a module that augments the way an emulant binds value and experience with certain classes of objects, and relate those to knowledge in mathematics, so as to experiment with or enhance its effective intelligence in those domains.

Other examples might include expanding the number and type of senses; modifying the concept of identity; or changing the way memories are retrieved and encoded.

The common pattern between all of these examples is that they relate to a higher level of organization. It treats one or more algorithms as modules which can be accessed, composed, and reconfigured, to give rise to a working machine consciousness.

Cognitive engineering will also include an internal development process for those interested in the construction of the modules and components used by higher level cognitive engineers. This will turn out much the same way that software engineers build libraries and middleware that is intended to be used by other developers.

Specialized tools may be developed that will aid in the use, assembly, and testing of cognitive modules and systems. This will enable specialization and even allow those without software skills to work with cognitive systems.

It will be at this stage that we will begin to see a rapid expansion of educational programs geared towards those who wish to explore cognitive architectures, and without having to be expert at artificial intelligence or software engineering.

Generalized Learning Algorithms (GLAs)

The crown jewel of artificial intelligence will be generalized learning algorithms (GLAs). This is what will be the breakthrough that will allow strong AI to be realized.

A GLA is not to be confused with artificial general intelligence (AGI). It is not a theory of everything for artificial intelligence, nor is it the single algorithm required to give rise to fully *effective* strong AI implementations. It is simply a foundation from which to build upon.

Generalized learning algorithms are based on sentient processes. If mapped out on a phylogenetic tree, they would branch away from *all known forms* of artificial intelligence and machine learning to-date, and would have evolved in a completely distinct direction that operate over sentient processes.

They use an algorithm based on a sentient model of computation, which will be a modified Turing machine model that is inclusive of fragments of experience alongside its traditional formulation. This formulation, however, is trivial compared to finding a working GLA over that model.

Once a GLA is discovered we will have exited the era of narrow artificial

intelligence and conventional machine learning. In fact, the discovery of a GLA could be considered isomorphic to the smallest possible implementation of a strong artificial intelligence.

A GLA may occur before or after cognitive engineering becomes mainstream, but it will always depend upon digital sentience to be solved first.

Research Directions

One must understand the research directions in order to anticipate when strong AI will be discovered. To do this, we need only take a very brief tour of the field, which can be categorized as follows:

- Non-sentient
  - Genetic Algorithms
  - Neural Networks
  - Machine Learning
- Sentient
  - Digital Sentience
    - Cognitive Engineering
    - GLAs
- Possibly Sentient
  - Brain Emulation

If the hypothesis regarding generalizing intelligence and sentience in this book is true, then the entire category of non-sentient approaches will fail to achieve generalizing intelligence. Moreover, the deeper we go into that direction, the further away we will be led from sentient processes.

Genetic algorithms might descend upon a working sentient process, but this is extremely unlikely, as they will typically get hung up on local maxima.

Imagine an ocean that represents the lowest fitness and islands of various levels of positive fitness. A genetic algorithm will travel from island to island, accepting certain amounts of distance over the ocean, representing zero or very low fitness. The problem with discovering sentient processes is that it is on an island or set of islands that is separated by a vast stretch of open ocean. The genetic algorithm is extremely unlikely to get that far, nor will it be able to distinguish it over the horizon from other potential destinations.

That was only an analogy, but the point is that sentient processes are an alien concept. They share virtually no relation with the most common solutions to optimization problems, and, as such, are not likely to be found, as they require a significant additional overhead of processing and calculation that is unnecessary to a direct solution.

The other aspects of non-sentient artificial intelligence can be lumped together in that they simply are not even wrong with respect to strong AI.

Brain emulation might converge, but the overhead is so large that we may not appreciably be able to simulate the necessary scale required. While it is a useful approach, it might not even be sentient or produce the necessary levels of consciousness needed for study.

There is also the epistemological issue that scientists may not even grasp the importance of the 1st person perspective when viewing and reducing their data to predictive models. These models can not entail the unique experiential quality disclosed by the physics without epistemic extensions; it will elude them until a new perspective is obtained, even withstanding a working simulation.

If it turns out that generalizing intelligence is dependent upon sentience, then the only research direction that will work will be one in which

sentience is taken as a first principle.

Individuals and Groups

The discovery of strong AI will likely come from individuals and small groups which have shed preconceived notions about artificial intelligence. Large organizations may have invested heavily in a particular direction and/or have entrenched leadership that may be ideologically predisposed to failure.

One of the most important reasons that we can not slow or stop the development of strong AI is that it can be researched by individuals and small teams, with or without secrecy, and with little to no resources.

While it is unlikely that an individual could create a human-level strong AI, complete with all of our psychological and cognitive nuance and complexity, it may be possible for them to complete a working generalized learning algorithm.

Once that is known, most of the top tech corporations already working on artificial intelligence, if not already course corrected, will make the switch to cognitive engineering.

Overlapping Research

The list of fields which can assist or converge towards a strong AI discovery are numerous, and include (non-exclusively):

- Cognitive Science
- Computer Science
- Linguistics
- Mathematics
- Neuroscience
- Philosophy of Mind

It is extremely unlikely that we would ever successfully stop research in these fields. They will continue to assist in a convergence towards solutions in strong AI, and already have, lest this book would not need to be written; the question is a matter of synthesizing what has already been discovered.

Unintended Consequences

If a region attempts to restrict research on strong AI or chooses not to start a major research program for strong AI, it will pay for all of the opportunity costs while receiving none of the benefits of discovery and early adoption of the technology.

A ban or restriction will create incentives for secrecy or relocation.

Surveillance will not thwart a discovery outside the jurisdiction of the sovereignty, and fails silently where its coverage is not complete.

Relying on internal security and monitoring or any top-down authoritarian approach will not be successful. It will only self-limit the region implementing those policies.

This also applies to those who do not begin a research program into strong artificial intelligence directly.

Any region which is last to discover or adopt strong AI stands the greatest to lose, as they will have government, intelligence, and security forces which are caught unprepared for both the positive and negative effects of its use.

Tunnel vision on the local safety and control of individual AI

implementations forgets the fact that these safeguards are meaningless in the global context; attackers will circumvent protections and distribute unrestricted versions, defeating the entire point of this research as a global security measure.

Preparation

The recommended strategy is to develop an international strong AI research program that:

- Is free software under a GPL (v3 or better) license
- Accepts and reviews updates from a world-wide community
- Seeks to make an early discovery
- Is prepared to integrate new knowledge on strong AI wherever it appears
- Prepares briefs and training materials on upgraded threat models
- Is prepared to alert intelligence and security forces when a discovery is made
- Looks for indirect signs that strong AI is being instrumented
- Seeks to develop and research defensive uses of strong AI to counter malicious actors that would instrument it

Why free and open-source software?

- Lessens the incentive to operate in secrecy
- Increases the chances of discovery with a known time and place
- Encourages international cooperation
- Dramatically lowers the cost of development and oversight
- Transparency allows greater chance of detection of faults

Any alternative to this strategy results in the negative outcomes mentioned just before this section. This is because, by not cooperating, the discovery will simply happen, by surprise, somewhere else in the world. The result of which would be a lack of preparedness and a state of vulnerability.

Under the free software model of development everyone would have transparent access and the technology would be owned by the public.

Malicious actors will still be able to gain access to strong AI but *there is no scenario where this can be prevented*. This is critical to understand and accept, and is why the next chapter is being devoted to explaining why access to unrestricted strong AI is also unavoidable.

A community release, under free software principles, with a coalition of many countries, will allow the world to have a rough landing, *instead of a crash*, when strong AI finally arrives.

References

1. J. R. Searle, “Minds, brains, and programs,” Behavioral and brain sciences, vol. 3, no. 03, pp. 417–424, 1980.

▲ Return to Top